Möglichkeiten und Grenzen von Big Data in der Forschung – Aktuelle Perspektiven Podiumsdiskussion der Jungen BWG am 4. September 2020

ALEXANDER WASZYNSKI, JANINA BAHNEMANN, SUSANA CASTILLO, PHILIPP KLAHN, PHILIPP OTTO und MARLIN ULMER

1. Ein neues Format

Im Rahmen des Gesellschaftsabends der Jungen BWG fand am 4. September 2020 und unter freiem Himmel eine klassenübergreifende, transdisziplinäre Podiumsdiskussion zum Thema „‚Big Data' als Ressource der Forschung – Aktuelle Perspektiven" statt. Die Ergebnisse dieser Diskussion werden hier in Form von drei gemeinsamen Thesen wiedergegeben.

Podium: Dr.-Ing. Susana Castillo (Institut für Computergraphik, TU Braunschweig), Jun.-Prof. Dr. Philipp Otto (Institut für Kartographie und Geoinformatik, LU Hannover) und Jun.-Prof. Dr. Marlin Ulmer (Institut für Wirtschaftsinformatik, TU Braunschweig).

Moderation: Dr. Alexander Waszynski (Seminar für Philosophie, TU Braunschweig).

Mit Diskussionsbeiträgen von Dr. Janina Bahnemann (Institut für Technische Chemie, LU Hannover), Prof. Dr. Nicole C. Karafyllis (Seminar für Philosophie, TU Braunschweig), Dr. Philipp Klahn (Institut für Organische Chemie, TU Braunschweig) und Jun.-Prof. Dr.-Ing. Ulrich Römer (Institut für Dynamik und Schwingungen, TU Braunschweig).

2. Einleitende Bemerkungen

Der Begriff Big Data ist im Zuge der Arbeit an frühen 3D-Visualisierungen geprägt und etwa zur Jahrtausendwende vonseiten der Wirtschaftsinformatik neu gefüllt und konzeptuell ausgearbeitet worden.[1] Die Minimaldefinition lautet: „*large* datasets that are produced in a *digital* form and can be analysed through *computational* tools." (Leonelli 2020) Es handelt sich klarerweise um große Datenmengen, aber diese zeichnen sich, so die gängige Einschätzung, dadurch aus, dass sie etablierte Verfahren der Datenverarbeitung überfordern. Als Kriterien werden neben den etablierten „three V's" – *volume*, *velocity,*[2] *variety* – mittlerweile auch genannt (vgl. Kitchin/McArdle 2016): *exhaustivity*,[3] *resolution*, *relationality*,[4] *scalability*[5] und *extensionality*. Vor allem *value* und *veracity* setzen sich neben den „three V's" vermehrt als Kriterien durch (vgl. Leonelli 2020). Big Data ist ein ‚mitwachsender' Begriff.

1 Anscheinend wird der Begriff erstmals in Anzeigen und Präsentationen der Firma *Silicon Graphics (SGI)* verwendet, darunter John Masheys „slide deck" mit dem Titel „Big Data and the Next Wave of InfraStress" (1998). Vgl. Diebold 2012.

2 Nennen lässt sich etwa die parallele Datenverarbeitung auf mehreren Servern, woraus sich extrem kurze Antwortzeiten und Möglichkeiten der Echtzeitprozessierung ergeben.

3 Es geht dabei um das Erfassen eines Ganzen (vgl. unten Abschnitt 2).

4 Gemeint ist Verbindung unterschiedlicher Datenbereiche untereinander (vgl. unten Abschnitt 2).

5 Dies betrifft die Möglichkeit, neue Datenbereiche zu implementieren (vgl. unten Abschnitt 2).

https://doi.org/10.24355/dbbs.084-202109030955-0

Zugleich ist nicht immer klar, was noch unter diesen Begriff fällt; ob es sich allein um ein informationstechnisches Problem handelt,[6] um eine epistemische Chance, ein forschungspolitisches Label oder um ein ohnehin alle wissenschaftlichen Akteure betreffendes Paradigma. Die Herausforderung ‚klassisch' gewordener Modelle des Weltverstehens durch rein datenbasierte Zugänge lässt sich in den Natur- und Ingenieurwissenschaften (Römer) ebenso beobachten wie in den Geisteswissenschaften (Waszynski).[7] Nicht selten wird mit Big Data die Hoffnung auf größere Objektivität verbunden. Im Jahr 2017 berichtete die Max-Planck-Gesellschaft, an ihrem Institut für Informatik in Saarbrücken sei es, in radikaler Umkehrung der Hypothesen-geleiteten Forschung, gelungen, „schon vorhandene Datensätze [zu] analysieren und daraus nachträglich Hypothesen und unerwartete Korrelationen [zu] extrahieren."[8] Doch auch dieser Ansatz bleibt auf Verfahren und Vorentscheidungen angewiesen, die zur Konstitution nutzbarer Daten wie zu deren Auswertung eingebracht werden müssen. Um Daten sammeln zu können, bedarf es der Annahmen darüber, was überhaupt als sammelns- und erhaltenswert zu erachten ist. Solche Annahmen lassen sich eigens reflektieren. So hat Sabina Leonelli in ihrem grundlegenden Artikel zu „Scientific Research and Big Data" (2020) aus wissenschaftsphilosophischer Perspektive betont: „Data need to be selected, cleaned and prepared to be subjected to statistical and computational analysis. The processes involved in separating data from noise, clustering data so that it is tractable, and integrating data of different formats turn out to be highly sophisticated and theoretically structured" (Leonelli 2020). Unsere Diskussion schließt an diesen Befund an und berührt damit zugleich das Programm des Querschnittsbereichs RECOLLECT der BWG.

3. Drei Thesen zum Umgang mit Big Data in der Forschung

1) Sammeln: Erheben – Speichern – Verarbeiten

Im Fachgebiet der Computergraphik bestimmen großen Datenmengen fast schon naturgemäß das Untersuchungsfeld. Allerdings besteht ein beträchtlicher Unterschied zwischen Daten, die in weitgehend unbearbeiteter Form vorliegen, und solchen, die bereits bearbeitet (z. B. retouchiert) worden sind (Castillo). Gerade bei visuellen Daten schließt die Erhebung das Sammeln von Gesammeltem ein, etwa im Falle der Arbeit mit Bilddatenbanken. Daten sind nicht, wie es der Wortsinn nahelegt, gegeben, sondern, was die Bedingungen ihrer Konstitution und ihres Erhalts betrifft, zugleich auch gemacht (Waszynski). Aus statistischer Sicht lässt sich die Komplexität von Daten als Spannung zwischen der Anzahl möglicher Charakteristika und der möglichen Beobachtungen beschreiben. Um valide Ergebnisse zu erhalten, ist eine methodologische Rahmung unabdingbar, die diese Spannung ausgleicht; eine entsprechend fokussierte Theoriearbeit kann es erlauben, auch auf Grundlage kleinerer Datensets belastbare Ergebnisse zu erzielen. Das schließt eine größere Effizienz hinsichtlich der zur Prozessierung und Speicherung benötigten Ressourcen ein (Otto). Zudem lassen sich so auch unter Bedingungen erschwerter Datenzugänglichkeit Forschungsthesen absichern. Die Auswertung bereits erhobener Daten, etwa zum Zweck von Prozessoptimierungen, wirft allerdings neue Fragestellungen auf, die in nachfolgende Erhebungen eingehen können, welche aber wiederum nicht allein einer bestimmten Zwecksetzung untergeordnet werden dürfen, sondern offen bleiben müssen für Unerwartetes (Ulmer). Das im Zuge einer Datenanalyse Ausgesonderte, der ‚Datenmüll', kann womöglich

6 Vgl. dazu die Aktivitäten im Rahmen des seit 2014 geförderten DFG-Schwerpunktprogramms 1736: *Algorithmen für große Datenmengen*, https://www.big-data-spp.de/ (zuletzt abgerufen am 12.11.2020).

7 Für die *Life Sciences* diskutiert dieses Paradigma ausführlich: Strasser 2019.

8 MaxPlanckForschung. Das Wissenschaftsmagazin der Max-Planck-Gesellschaft, I.2017, S. 4. Abrufbar unter https://www.mpg.de/11248402/MPF_2017_1.pdf (zuletzt abgerufen am 12.11.2020).

https://doi.org/10.24355/dbbs.084-202109030955-0

später und unter neuen Fragestellungen an Relevanz gewinnen, was Fragen nach einer Vorratsspeicherung oder sogar Datenarchäologie[9] aufwirft (Waszynski). Das bloße Speichern ist dabei eine Ressourcenfrage; kritischer ist es, Zugänglichkeit über Zeit zu gewährleisten. Damit das Material auffindbar und nutzbar bleibt, müssen Indizes, Modi und Gründe der Ersterfassung mitgespeichert werden (Otto). Allerdings werden z.B. im Bereich der theoretischen Chemie, etwa beim molekularen Docking, schnell Grenzen der Rechenleistung erreicht; mitunter ist zur Validierung ein rascherer Praxistest zielführender als die bloße Speicherung immenser Datenmengen auf Verdacht (Klahn).

These 1: Die Parameter des Sammelns und ‚Entsammelns'[10] von Daten müssen transparent und nachprüfbar bleiben. Um Effizienz und Plausibilität zu gewährleisten, bedarf es einer Intensivierung der Arbeit am theoretischen und methodologischen Rahmenwerk datenbasierter Forschung. Das schließt eine Problematisierung des Begriffs Big Data ein.

2) Grenzen: Definitionen – Ansprüche – Praxis

Die Grenzen von Big Data lassen sich aus verschiedenen Blickwinkeln beleuchten. In definitorischer Hinsicht birgt eine über die „three V's" gefundene Bestimmung die Schwierigkeit, angesichts rapider informationstechnischer Entwicklungsschritte eine zeitabhängige Definition zu erhalten (Otto). Daher lässt sich fragen, ob der Begriff Big Data nicht eher als flexible Problemmarkierung und weniger als fester Terminus verstanden werden muss (Waszynski). Um eine große Bandbreite unterschiedlicher Datentypen (*variety*) auswerten zu können, bedarf es eines gezielten Datenmanagements,[11] das die Schnittstellen bearbeitet. Dies gilt bereits für den Umgang mit ‚medium data', etwa im Falle von Transportwegeoptimierungen (Ulmer). Auch was die Datenformate betrifft, bleibt das Material letztlich heterogen. Ein internationaler Standard für alle Datentypen, der diese Heterogenität aufheben würde, ist praktisch nicht umsetzbar (Castillo). Um eine ergebnisorientierte Auswertung großer Datenmengen zu gewährleisten, ist offenbar ein gezieltes ‚blackboxing' bzw. das Delegieren von Zuständigkeiten (vgl. Abschnitt 3) notwendig (Waszynski). Big Data lässt sich auch als ein *Anspruch* beschreiben, nicht zuletzt auf globale Gültigkeit. Bisweilen schlagen sich, wie in Wahrnehmung und Interpretation des Gesichtsausdrucks, regionale Unterschiede in Datensätzen deutlich nieder; umgekehrt werden durch Annahmen über solche Unterschiede Ergebnisse möglicherweise präformiert. Um dies im Detail herauszuarbeiten, bedürfte es der globalen Verknüpfung aller Teilergebnisse und Einzeldaten (Castillo). In pragmatischer Hinsicht ist außerdem zu überlegen, ob sich Fragestellungen, die scheinbar einen Big Data-Zugang erforderlich machen, nicht auch mit kleineren Skalierungen erfolgreich bearbeiten lassen (vgl. Abschnitt 1). Große Datensätze können zwar

9 Der Begriff der Datenarchäologie hat eine Vorgeschichte in der Ozeanographie der 1990er Jahre. In einem internationalen Forschungsprojekt („GODAR") sollten nicht nur bestehende Aufzeichnungen digitalisiert, sondern auch, was sammlungstheoretisch nicht minder interessant ist, bereits in digitaler Form vorhandene Daten vor dem Verlust bewahrt werden. Zu den 1994 genannten Faktoren zählen: „media degradation such as fading ink or magnetic fields", „environmental catastrophes such as fires and floods", „simple neglect", „the retirement of individuals who know how to access these data or know the metadata associated with these data that make them useable to other scientists" (Levitus et al. 1994, 2).

10 ‚Entsammeln' ist ein in der Museologie geläufiger Terminus, der unterschiedliche Verfahren des Aussonderns erfasst und sich auf den 1972 von John Canaday geprägten Begriff der „deaccession" zurückführen lässt.

11 Aus wirtschaftsinformatischer Perspektive ist der Stellenwert einer kritischen „Data Gouvernance" betont worden, die der Diversität und dem Schutz von Daten gleichermaßen Rechnung trägt (Buhl et al. 2013).

https://doi.org/10.24355/dbbs.084-202109030955-0

Ergebnisse produzieren, verleiten aber mitunter zu Schlussfolgerungen, die bei näherer Betrachtung irreführend sein können, wenn die theoretischen Prämissen nicht richtig gewählt worden sind. Vor allem im medizinischen Bereich kann dies schnell problematisch werden (Otto). Es bedarf eines geschulten menschlichen Blicks auf die Daten. Zwar gibt es Beispiele für Datenprozessierungen, etwa den Börsenhandel, die auch ohne eine humane Intervention in sich automatisiert funktionieren (Ulmer); in Rücksicht auf die Komplexität nicht nur der Datenerhebung und -prozessierung, sondern auch auf deren mögliche Folgewirkungen lässt sich aber der praktische Erfolg nicht zum alleinigen Kriterium dafür machen, ob es einer zwischengeschalteten, Werte und Auswertungen überprüfenden Instanz bedarf (Waszynski).

These 2: Der Begriff Big Data umfasst nicht nur aufgrund ihrer Größe, Zusammensetzung und Verarbeitbarkeit besonders ausgezeichnete Datenmengen, sondern schließt auch eine gewisse Unschärfe ein: einerseits in Bezug auf seine Zeitabhängigkeit, andererseits in Bezug auf die innere Kohärenz dieser Datenmengen. Das als Big Data Verstandene ist, noch vor seiner Auswertbarkeit, Gegenstand bearbeitender Eingriffe. Als Begrenzung der Reichweite des Begriffs lässt sich zudem sein Anspruchscharakter nennen.

3) Interdisziplinarität: Ausblicke

Es besteht Konsens darüber, dass im Umgang mit Big Data disziplinäre Grenzen und Zuständigkeitsbereiche immer wieder überschritten werden müssen. Damit die Kriterien *value* and *veracity* einlösbar sind, müssen mitunter Expert*innen aus ganz anderen Fachbereichen hinzugezogen werden, etwa aus der Psychologie, wenn es um das auf visuellen Daten basierende Studium der Feinheiten des Gesichtsausdrucks geht (Castillo). Es ist ebenfalls Konsens, dass ethische und wissenschaftstheoretische Erwägungen in die Forschungsarbeit einbezogen werden müssen. Big Data bietet eine Chance, über historisch gewachsene Privilegierungen und Geschlechterzuschreibungen hinauszugehen (Ulmer). Der Umgang mit Daten bleibt dennoch, z.B. bei genomabhängigen Entscheidungen im Bereich der Medizin, anfällig für Diskriminierung (Klahn). Bereits in die Erhebung von Daten gehen normative Vorentscheidungen ein, nicht erst in die Auswertung. Dies hat die *Gender Medizin* in den letzten Jahrzehnten deutlich herausgearbeitet. Big Data benötigt Verfahren der Standardisierung (vgl. Abschnitt 2), kluge Entscheidungen brauchen jedoch Diversifizierung. Als Beschleunigungsstrategie ist Big Data dort angebracht, wo Prozesse beschleunigt werden müssen, was nicht immer der Fall ist (Karafyllis). In der Tat gibt es Fälle, in denen bereits die Sensoren diskriminieren (Ulmer). Am Beispiel des impliziten Imperativs zur möglichst breiten Datenerfassung in anwendungsorientierten Forschungskontexten der Chemie und Biochemie lässt sich die Notwendigkeit zur interdisziplinären Zusammenarbeit deutlich demonstrieren; die Netzwerkbildung wird damit zu einem integralen Bestandteil gegenwärtigen wissenschaftlichen Arbeitens (Bahnemann).

These 3: In Anbetracht des Big Data inhärenten Prinzips der Delegation ist es unabdingbar, dass Kollaborationen weiter etabliert und verstetigt werden. Das bedeutet auch, wissenschaftstheoretische und ethische Erwägungen bereits in das Forschungsdesign einzubeziehen.

Literatur

BUHL, H.U. et al. (2013): Big Data – Ein (ir-)relevanter Modebegriff für Wissenschaft und Praxis?, in: Wirtschaftsinformatik & Management **5**, S. 24–31. DOI: 10.1365/s35764-013-0275-6.

https://doi.org/10.24355/dbbs.084-202109030955-0

DIEBOLD, F.X. (2012): On the Origin(s) and Development of the Term ‘Big Data’, September 21, 2012, *PIER Working Paper* No. 12–037. DOI: 10.2139/ssrn.2152421.
KITCHIN, R. & McARDLE, G. (2016): What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets, in: Big Data & Society, January–June 2016, S. 1–10. DOI: 10.1177/2053951716631130.
LEONELLI, S. (2020): Scientific Research and Big Data, in: Zalta, E. N. (Hg.), The Stanford Encyclopedia of Philosophy (Summer 2020 Edition). URL: https://plato.stanford.edu/archives/sum2020/entries/science-big-data/ (zuletzt abgerufen am 12.11.2020).
LEVITUS, S. et al. (1994): Results of the NODC Oceanographic Data Archaeology and Rescue Projects. Report 1. Key to Oceanographic Records Documentation No. 19, NODC, Washington, D.C. Abrufbar unter: ftp://ftp.library.noaa.gov/noaa_documents.lib/NESDIS/NODC/key_to_oceanographic_records/no_19.pdf (zuletzt 26.11.2020).
STRASSER, B. J. (2019): Collecting Experiments: Making Big Data Biology. Chicago.

https://doi.org/10.24355/dbbs.084-202109030955-0